科学通报 2026 年 第 71 卷 第 30 期: 6885 ~ 6902
观 点
《中国科学》杂志社
SCIENCE CHINA PRESS

热管理学科前沿: 从微纳传热到AI算力基础设施

杨荣贵[1*], 李小波[2*]

1. 北京大学工学院能源与资源工程系, 北京 100871
2. 华中科技大学能源与动力工程学院, 武汉 430074
* 联系人, E-mail: ronggui@pku.edu.cn; xbli35@hust.edu.cn

当智能手机连续运行三局游戏后轻微发烫, 当AI芯片总功率突破1200 W、接近家用电磁炉水平, 当数据中心功率密度从每机架10 kW飙升至250 kW, 这些日常场景与技术现象的背后, 实则是同一物理问题在不同尺度下的呈现, 即热管理问题.

热管理是对系统、设备或器件内部热量的产生、传递、分布与耗散过程开展主动设计、精确控制与系统优化, 以确保其工作温度始终维持在安全、高效、可靠区间的一整套技术与方案体系. 从智能手机、AI芯片到数据中心、新能源汽车乃至航空航天装备, 热管理已深度渗透至现代科技的各个战略领域.

然而, 热管理正面临前所未有的挑战. 单颗AI芯片的局部热流密度已超过300 W/cm^2, 数倍于典型核反应堆堆芯的热流密度; 传统风冷技术的极限散热热流密度约为100 W/cm^2, 在高热流密度、高密度机架场景下, 液冷成为突破散热限制的主流技术路线. 更深刻的挑战在于尺度跨越: 从纳米尺度下声子的量子输运, 到微米级通道内气泡的成核沸腾, 再到米级机柜的气流组织与千公里级卫星星座的热控调度, 热管理需在十余个数量级的尺度范围内建立统一的热流秩序. 这一目标的实现, 需要凝聚态物理、微电子学、流体力学、传热学、材料科学等多学科的深度交叉与协同创新.

2022年, 中国科学院发布《中国学科发展战略·电子设备热管理》[1], 标志着热管理领域已从工程技术配套环节, 跃升为国家战略科技力量必须系统布局的基础研究领域. 作为新质生产力的关键基础支撑, 热管理正经历从"配套环节"到"核心约束"的身份转变. 其每一次技术突破, 都将为AI算力提升、清洁能源开发、商业航天发展等战略领域拓展新的可能性.

本文旨在为读者描绘热管理的学科地图, 依次阐述热管理的学科边界与知识体系, 梳理学科建制化发展历程, 解析声子输运、界面热阻、尺度跨越三大基础物理问题, 聚焦集成电路领域, 剖析功耗墙、三维集成技术、材料革新与方法论创新等核心议题, 针对AI数据中心阐述功率密度提升趋势与液冷技术体系, 探讨热管理作为新质生产力基础的战略价值跃迁, 揭示学科发展面临的核心挑战, 并对未来研究方向进行展望.



杨荣贵 北京大学工学院讲席教授、能源与资源工程系主任. 曾任科罗拉多大学终身正教授、华中科技大学二级教授. 长期从事微纳米尺度热输运、电子器件热管理、光热界面与辐射制冷、高效相变传热以及热储能研究. 研究成果获MIT Technology Review's TR35 Award、DARPA Young Faculty Award及NSF CAREER Award等. 团队辐射制冷超材料薄膜研究入选《物理世界》"2017年度全球十大物理突破". 主持中、美国家级项目40余项.



李小波 华中科技大学能源与动力工程学院副教授. 研究方向为热储能与电子器件热管理. 2004年获中国科学技术大学学士学位, 2007年获中国科学院工程热物理研究所硕士学位, 2012年获科罗拉多大学博尔德分校博士学位, 2012~2014年于麻省理工学院从事博士后研究. 入选第十三批"3551光谷人才计划".

引用格式: 杨荣贵, 李小波. 热管理学科前沿: 从微纳传热到AI算力基础设施. 科学通报, 2026, 71: 6885–6902
Yang R-G, Li X-B. Advances in thermal management: from micro/nanoscale heat transfer to AI computing infrastructure (in Chinese). Chin Sci Bull, 2026, 71: 6885–6902, doi: 10.1360/CSB-2026-0245

© 2026《中国科学》杂志社

1 热管理是一门什么样的学科?

1.1 定义与边界: 从“配套”到“核心”

《中国学科发展战略·电子设备热管理》对“电子设备热管理”的定义是: 研究电子元器件及设备在工作过程中的热产生、热传输、热耗散规律, 并基于这些规律发展热控制方法与技术的交叉学科[1]. 该定义包含若干核心概念, 值得深入阐释.

第一, “热产生”: 热并非芯片设计完成后才出现的衍生问题, 而是电子输运的必然伴生物. 当电子在导体中流动并与晶格发生碰撞, 当晶体管进行开关操作并对电容充放电, 当高频信号在介质中传播并引发介质极化损耗时, 能量耗散是物理定律所施加的必然代价, 任何计算过程均无法规避这一规律[2]. 热管理的首要任务, 便是理解该代价的产生根源.

第二, “热传输”: 热量自芯片内部纳米级的热点(部分热点宽度仅为几十个原子)出发, 依次经过晶体管层、金属互连层、衬底、封装热界面材料、散热盖以及外部散热器, 最终传递至环境空气或冷却液, 这是一条跨越十余个数量级的传输路径[3]. 在该路径上, 导热、对流、辐射三种热传递机制交替发挥作用, 每种机制均存在其物理极限, 且均会受到界面因素的制约.

第三, “规律”: 热管理并非试错的艺术, 而是建立在傅里叶导热定律、牛顿对流冷却定律、斯蒂芬-玻尔兹曼热辐射四次方定律、汽液相变热力学等基础物理之上的工程科学[3,4]. 高中及大学物理课程中的经典公式(如傅里叶定律、牛顿冷却定律), 至今仍是热管理工程师日常分析的核心工具.

第四, “交叉”: 热管理的学科边界具有模糊性, 这正是其学科特色所在. 它向下深入固体物理(涵盖能量与信息载流子的波粒二象性、电子与声子输运、电子-声子耦合等内容)、流体力学(包括单相/两相流动、微尺度流动等方向)、热力学领域及材料科学领域(涉及高导热复合材料、热界面材料等研究); 向上可衔接电子工程领域(如集成电路设计、先进封装技术)与计算机科学领域(如动态热管理、温度感知调度方法); 向外可延伸至能源工程领域(例如制冷与超低温技术、核聚变堆热控系统)、航天工程领域(比如航天器热控技术、热防护系统)、机器人领域及新能源交通领域(聚焦电池、电机、电控“三电”系统的热管理问题). 在很多院系和学科门类都有热管理的影子[1].

1.2 学科树: 热管理在学什么?

依据《中国学科发展战略·电子设备热管理》的学科梳理[1], 热管理的知识体系可划分为五个层次, 每个层次均回应一个根本性问题.

(1) 基础机理层面: 热从哪里来? 热往哪里走? 这是热管理作为“科学”的根基. 核心问题涵盖: (a) 芯片产热机理. 在互补金属氧化物半导体(complementary metal-oxide-semiconductor, CMOS)晶体管开关过程中, 焦耳热于纳米尺度沟道内的分布情况如何? 电子-声子散射怎样转化为晶格温升[2]? (b) 微纳热输运. 当特征尺寸小于声子平均自由程(室温下在硅中约为300 nm)时, 傅里叶定律是否依然成立? 若不成立, 如何描述热流[4~6]? (c) 界面热输运. 热量流经两种不同材料的界面时, 为何会发生“反射”现象? 如何对界面热阻(卡皮查热阻)进行预测与测量[7]? (d) 电-热-力多场耦合. 温度梯度如何诱发热应力? 封装翘曲又如何反向增大接触热阻?

(2) 核心方法层: 如何把热带走? 这是热管理作为“技术”的主体, 具体包括: (a) 芯片级热管理, 如纳米导热结构、近结点微通道、嵌入式流动与相变、背面蚀刻微流道; (b) 封装级热扩展, 如高导热复合材料(金刚石/铜、碳化硅)、蒸汽腔均温板(vapor chamber, VC)、热管[8]; (c) 界面热管理, 如热界面材料(thermal interface material, TIM), 包括芯片与散热盖间的一级界面材料(thermal interface material 1, TIM1)、散热盖与散热器间的二级界面材料(thermal interface material 2, TIM2), 液态金属、石墨烯导热膜[9]; (d) 高效散热技术, 如单相/两相液冷冷板、喷雾冷却、射流冲击、浸没冷却[10,11].

(3) 系统集成层: 如何实现热系统与电系统的协同工作? 从器件到设备层面, 热管理需与电气设计、结构设计及软件调度协同推进. 具体涵盖: (a) 热设计方法, 如计算流体动力学(computational fluid dynamics, CFD)数值仿真、多物理场耦合建模、降阶模型; (b) 电力电子热管理, 如绝缘栅双极型晶体管(insulated gate bipolar transistor, IGBT)、碳化硅(silicon carbide, SiC)和氮化镓(gallium nitride, GaN)宽禁带器件、逆变器; (c) 数据中心热管理, 如气流组织优化、冷热通道隔离、液冷系统集成、废热回收利用[12,13].

(4) 软件定义层: 如何用算法管理热? 这是近十年来兴起的新方向. 热管理正在被“软件化”, 具体涵盖: 温度场高精度感知与重构(从稀疏传感器到全场估计); 功耗-温度关联建模(从物理模型到数据驱动); 任务调度与热管理协同优化(负载均衡、动态电压频率调整)[14].

(5) 交叉应用层: 热管理还能拓展至哪些领域? 热管理从电子设备起步, 正朝着更广阔的领域延伸, 这正是其作为“新质生产力”关键基础的核心原因. 具体涵盖: (a) 集成电路, 如3D异构集成、芯粒(chiplet)、晶体管级热设计[15,16]; (b) AI数据中心, 如冷板、浸没液冷、两相冷却、余热利用[10,12]; (c) 新能源汽车与低空飞行器, 如动力电池热管理、电机与电控集成热控[17]; (d) 机器人, 如芯片热管理、电池热管理、电机与电控集成热控; (e) 量子计算, 如稀释制冷、极低温高真空环境; (f) 商业航天, 如卫星热控、空间站两相流、深空探测器辐射冷却; (g) 聚变能源, 如第一壁和偏滤器冷却、聚变能的输出与发电.

这是一棵仍处于生长状态的学科树, 如图1所示. 图1以“根-干-枝-叶”结构呈现该知识体系: 根基对应基础物理, 主

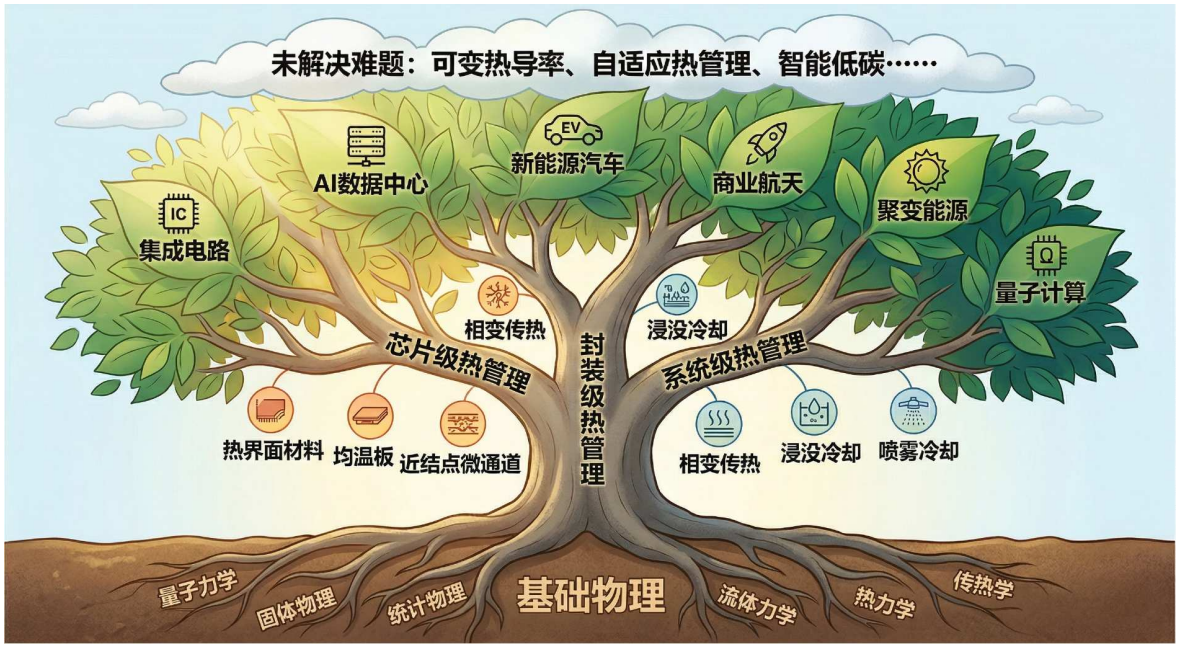


图 1　热管理学科知识树：热管理的学科体系呈现为“根-干-枝-叶”结构．根基是基础物理学，涵盖量子力学、固体物理、统计物理、流体力学、热力学、传热学；主干为核心方法，包括芯片级、封装级、系统级热管理技术，并延伸出具体技术，包括近结点微通道、均温板、热界面材料、浸没冷却、喷雾冷却等．叶片为交叉应用领域，涉及集成电路、AI数据中心、机器人、新能源汽车、商业航天、聚变能源、量子计算．树冠的“未解决难题”象征该学科的开放性．本图基于作者原创思路，采用AI工具绘制

Figure 1　Disciplinary knowledge tree of thermal management. The disciplinary system of thermal management presents a “root-trunk-branch-leaf” structure. The root is fundamental physics, including quantum mechanics, solid state physics, statistical physics, fluid mechanics, thermodynamics, and heat transfer; the trunk represents core methods, including chip-level, package-level, and system-level thermal management technologies, extending to specific technologies such as near-junction microchannels, vapor chambers, thermal interface materials, immersion cooling, and spray cooling; the leaves stand for cross-cutting application fields, involving integrated circuits, AI data centers, robots, new energy vehicles, commercial aerospace, fusion energy, and quantum computing; the “unsolved problems” in the crown symbolize the openness of the discipline. This figure is based on the author’s original idea and is drawn using AI tools

干为核心方法，并延伸出具体技术，叶片为交叉应用领域，树冠的“未解决难题”则象征该学科各分支均存在尚未解决的重要问题(部分已逾五十年，部分则于近年才被界定)．

2　学科简史：热管理如何成为一门独立学科？

2.1　前史：从蒸汽机到晶体管

热科学与工程的学科发展历程，本质上是一段“尺度向内迁移”的演进史．2014年，在国际传热大会的傅里叶讲座上，美国马里兰大学教授、美国国防高级研究计划局(Defense Advanced Research Projects Agency，DARPA)微系统技术办公室项目主管阿夫拉姆·巴尔-科恩(Avi Bar-Cohen)作了题为“热科学与工程——两百年间从宏观到纳米”的综述报告[18]．他提出，19世纪为蒸汽机研发提供支撑的传热学原理，当前可能被错误地广泛运用于纳米电子器件的设计与优化．此论断揭示了热管理学科的一个深层次特性：其基本现象虽未发生改变，然而尺度却跨越了十余个数量级，因而所有问题均需重新解答．

1965年，戈登·摩尔(Gordon Moore)提出其著名的经验预测时，芯片的功耗密度约处于1 W/cm^{2}量级，采用一枚简易散热片即可满足散热需求．工程师将热设计置于封装手册的最后一章，作为“必要时参考”的附录内容．20世纪70~80年代，集成电路的集成度依据摩尔定律稳步提升，功耗密度上升至10 W/cm^2．风扇开始被应用于机箱之中，铝挤型散热片与轴流风扇的组合成为个人计算机(personal computer，PC)工业的标准配置．然而，在此阶段，热设计仍被视作“后端工程”，即芯片设计完成、封装选型确定之后，再由散热器厂商提供相应的解决方案．

2.2　建制化：DARPA三十年的路线图

热管理在全球范围内的学科建制化进程并非同步．美国以DARPA专项为标志，率先完成从“配套环节”向“独立学科”的身份转换；中国虽起步稍晚，但凭借应用场景的规模化优势，在国家重大需求牵引下实现了快速跟进与局部引领．以下以DARPA三十年路线图为参照，同步勾画中国热管理学科建制化的关键节点．

1991年, DARPA启动了“金刚石3D多芯片模块”计划. 这是有文献可考的、首个由国家科研资助机构设立的、以热管理为核心目标的专项研究[19]. 此计划的背景值得深入剖析. 当时, 高性能计算芯片的性能提升开始受到片上互连延迟的严重限制——电子在导线中的传输速度是有限的, 芯片尺寸越大, 信号从一端传输至另一端所需的时间越长. 一个合理的解决方案是: 将多个芯片堆叠成三维结构, 用垂直方向的短互连替代水平方向的长互连, 从而大幅缩短信号路径.

然而, 三维集成引发了全新的问题: 如何处理热量? 在平面芯片中, 热量能够向下通过衬底、向上通过封装盖板进行双向扩散; 在三维堆叠结构中, 下层芯片被上层芯片完全覆盖, 热量仅能从边缘极为有限的路径排出, 从而形成严重的“热阱”效应. DARPA的解决策略是: 采用人造金刚石作为热扩展衬底. 金刚石的热导率超过2000 W/(m K), 是铜的五倍、铝的近十倍. 理论计算显示, 1 mm厚的金刚石衬底能够支持400 W的芯片功率, 并将温升控制在25℃以内.

这一决策确立了此后三十年热管理研究的一个根本范式: 热不再是芯片性能提升后的被动补救措施, 而是与芯片架构同步设计的前置约束. 从1990年代到2020年代, DARPA持续布局热管理专项, 见表1和图2[19]. 表1将三十年路线划分为可行性验证、技术拆解、协同设计、新尺度新场景四个阶段, 图2以时间轴形式可视化这一演进脉络, 二者共同反映了该学科从“被动应对”到“主动储备”的战略转型.

这份跨度达三十年的路线图揭示出一个事实: 热管理领域的每一次进阶, 并非被动的“问题应对”, 而是主动的“能力储备”[19]. 当商用芯片功耗仅为几十瓦时, DARPA便已开展400 W金刚石散热方案的研究; 当液冷技术尚未在数据中心广泛普及之时, DARPA已着手探索微尺度两相流热管理技术; 当三维集成技术刚刚进入产业视野之际, DARPA已开始研发晶体管级的热仿真工具. 这正是热管理作为“战略学科”的真正内涵: 它并非在问题爆发后才采取补救措施, 而是提前十年对物理约束进行预判, 在问题演变为发展瓶颈之前储备相应的解决方案.

2.3 学科成熟: 国家战略与中国路径

2022年, 中国科学院把电子设备热管理纳入《中国学科发展战略》丛书[1]. 此乃热管理在中国学术建制化进程中的标志性事件. 该丛书定位为“国家科学思想库·学术引领系列”, 其选题标准极为严格, 要求学科需具备一定研究基础, 处于快速发展阶段, 且对国家重大战略需求有实质性支撑. 与之并列的选题包括量子光学、合成生物学、深空探测、海洋工程. 在该书最后一章, 专门探讨了学科建设与人才培养问题. 其中有一段话值得全文引用[1]: “电子设备热管理学科正处于从‘跟随型’向‘引领型’转变的关键时期. 学科发展的必然性、专业人才的缺乏、巨大的市场需求三者形成了尖锐矛盾. 我国电子设备热管理领域面临的人才缺口, 已从‘储备性短缺’转变为‘结构性紧缺’. ”换言之, 该领域的前沿问题与日俱增, 而能够独立攻关这些难题的复合型人才却严重不足. 简言之, “课题等人做, 但会做的人不够”. 这一判断并非夸大其词. 2025年IDTechEx的报告显示, 全球数据中心冷却市场规模预计将在未来十年内增长至数千亿人民币(数百亿美元)的规模. 其中, 仅直接芯片液冷细分市场, 预计将从2025

表 1 美国DARPA热管理三十年路线

Table 1 The 30-year roadmap of thermal management by DARPA, USA

阶段	项目	项目内容
1990年代 可行性验证期	1991年 金刚石3D MCM	验证高导热衬底在三维集成中的可行性
	1998年 HERETIC	探索固态/流体制冷技术与高密度电子器件的集成
2000年代 技术拆解期	2005年 MCC(微尺度低温制冷器)	将制冷功能微缩至芯片尺度
	2008~2012年 TMT(热管理技术)	将芯片封装的完整热阻链拆解为五个环节: 芯片热基底、微型空冷换热器、纳米热界面、主动冷却模块、近结点热输运, 分别设立专项进行技术攻关
2010年代 协同设计期	2012年 ICECool基础(芯片内/芯片间增强冷却)	提出电-热-力协同设计方法论, 在二维/三维拓扑架构下同步优化电气性能与热流路径
	2013年 ICECool应用	探索单相液冷与蒸发冷却在芯片封装的集成
2020年代 新尺度、新场景期	2023年 Thermonat	从单个CMOS晶体管到数十万晶体管电路的多尺度电-热耦合建模
	2023年 Minitherms3D	面向三维异质集成的微型热管理系统
	2023年 THREADS	射频器件晶体管级热管理

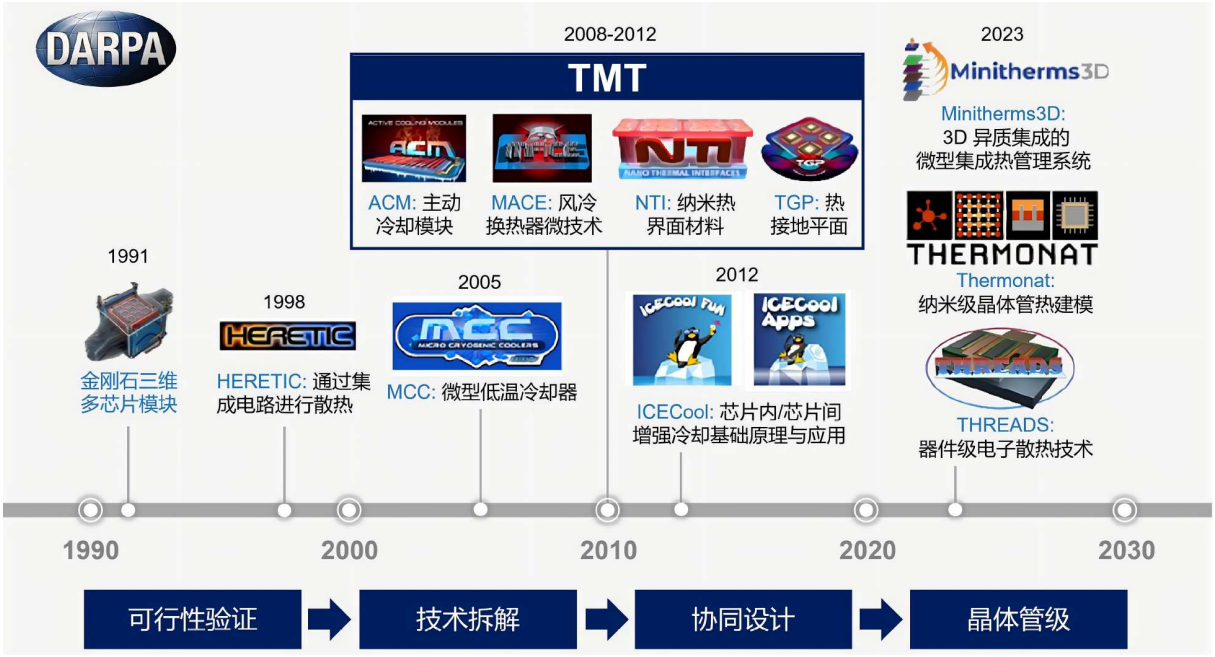


图 2 美国DARPA热管理三十年路线图(1991~2023). 该时间轴系统呈现了DARPA热管理专项的演进脉络. 1991年, “金刚石3D MCM”项目首次将热管理确立为独立研究方向; 2008~2012年间, “TMT”项目将热阻链拆解为五个关键环节进行专项攻关; 2012年, “ICECool”项目开创了电-热-力协同设计的新范式; 2023年, DARPA集中启动Thermonat、Minitherms3D、THREADS三大项目, 将研究尺度推进至晶体管级. 这一路线图清晰揭示了热管理从“被动应对”向“主动储备”的战略转型. 本图采用AI工具绘制

Figure 2 The 30-year roadmap of DARPA thermal management programs (1991–2023). This timeline systematically presents the evolutionary context of DARPA's thermal management programs. In 1991, the “Diamond 3D MCM” program established thermal management as an independent research direction for the first time. From 2008 to 2012, the “TMT” program deconstructed the thermal resistance chain into five key links for targeted research. In 2012, the “ICECool” program pioneered a new paradigm of electrical-thermal-mechanical collaborative design. In 2023, DARPA launched three major projects: Thermonat, Minitherms3D, and THREADS, pushing the research scale to the transistor level. This roadmap clearly reveals the strategic transformation of thermal management from “passive response” to “active reserve”. This figure is drawn using AI tools

年的340~680亿元人民币(50~100亿美元)增长至2035年的超过3400亿元人民币(510亿美元). 与此同时, 具备跨学科背景(传热学、流体力学、材料学、电子工程)的热管理工程师的供需缺口正以每年30%的速度持续扩大[12].

中国科学院2006年启动“微纳制造与系统集成”先导专项, 热管理作为关键支撑技术纳入布局; 2010年前后, 华为、中兴等企业开始建立芯片热设计能力, 标志中国产业端需求的觉醒; 2017年, 科技部“十三五”国家重点研发计划“变革性技术关键科学问题”重点专项设立“高效热管理”方向; 2018年, 中国科学院启动“电子信息材料与器件”战略性先导科技专项, 三维集成热管理为核心攻关内容. 相较于DARPA三十年连续专项的系统迭代, 国内研究布局呈现多部门协同推进的特点, 长期性、贯通性的顶层路线图仍在完善中. 然而, 中国拥有全球最大的消费电子制造、新能源汽车和数据中心增量市场, 终端设备制造商与算力基础设施运营商已具备从芯片封装到系统集成的全链条工程化能力, 这种“需求牵引+场景驱动”的产业优势, 正反向推动基础研究与技术标准的快速迭代.

3 学科根基: 热管理的“第一性原理”

在开启集成电路与人工智能数据中心相关研究的宏大篇章之前, 我们需先解答一个更为基础的问题: 热管理的物理根基究竟是什么? 这并非所谓的“背景知识”, 而是每位热管理工程师在日常工作中都必须直面的现实挑战.

3.1 声子: 热的量子载体

在高中与大学物理范畴内, 热传导被界定为“分子热运动的传递”. 此概念于宏观尺度下具有适用性, 但在纳米尺度下则完全失效. 声子(phonon)作为晶格振动波的量子化能量子, 是半导体材料中热量的主要承载者[2]. 可将其形象地理解为“晶格的集体振动模式”, 即每个原子均在自身平衡位置附近进行振动; 然而由于原子间存在化学键, 相邻原子的振动并非相互独立, 而是以波的形式于晶格中传播, 具体见图3.

声子的物理世界与电子的物理世界存在显著差异, 具体表现为: (1) 声子虽无质量, 却具备动量、能量以及平均自由程. 在纯净的硅单晶内, 在室温条件下, 声子的平均自由程约为300 nm, 即声子每行进300 nm便会与另一个声子或其他缺

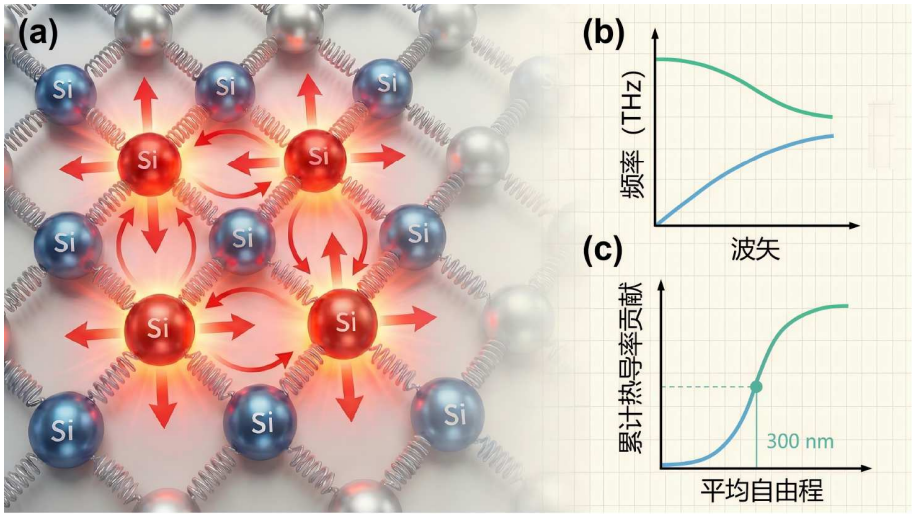


图 3 (a) 纳米尺度热载体——声子的输运示意图. (b) 硅晶体中声子色散关系示意图, 其中横轴为波矢, 纵轴为频率, 不同分支代表不同的振动模式. (c) 声子平均自由程累积分布, 表明硅中声子的平均自由程约为300 nm, 且具有宽广的频谱分布. 当器件特征尺寸小于平均自由程时, 声子呈现弹道输运, 傅里叶定律不再适用, 需求解玻尔兹曼输运方程. 本图基于作者原创思路, 采用AI工具绘制

Figure 3 (a) Schematic diagram of phonon transport as nanoscale heat carrier. (b) Schematic of phonon dispersion relations in silicon, where the horizontal axis represents the wave vector and the vertical axis represents the frequency; different branches correspond to different vibration modes. (c) Cumulative distribution of phonon mean free paths, indicating that the mean free path of phonons in silicon is approximately 300 nanometers with a broad spectrum distribution. When the characteristic size of the device is smaller than the mean free path, phonons exhibit ballistic transport and Fourier's law is no longer applicable, requiring the solution of the Boltzmann transport equation. This figure is based on the author's original idea and is drawn using AI tools

陷产生一次碰撞. (2) 声子频谱范围宽广, 涵盖从低频声学支至高频光学支, 各频率声子所携带能量、群速度以及遭遇的散射概率均存在差异. (3) 声子的输运呈现半经典特性. 当器件的特征尺寸大于声子平均自由程时, 热传导过程能够运用傅里叶定律进行描述; 而当特征尺寸小于平均自由程时, 声子的弹道飞行现象致使傅里叶定律不再适用, 此时必须求解玻尔兹曼输运方程[20,21].

为何这对集成电路热管理具有至关重要的意义? 原因在于, 当今先进制程的晶体管, 其关键尺寸已小于声子平均自由程. 3 nm节点的鳍片宽度以及纳米片厚度均处于10 nm量级; 这表明, 在晶体管工作的核心区域, 我们甚至无法准确回答"热导率是多少"这一最基本的问题. 这是由于热导率本身属于宏观统计平均概念, 在纳米尺度下已失去定义[20~22].

热管理领域的基础研究前沿聚焦于以下核心问题: 如何在非傅里叶区域实现温度场的精准预测? 如何基于第一性原理开展声子-电子耦合的计算研究[20,23]? 如何运用量子输运理论对热耗散过程进行科学描述? 这些问题并非理论家书斋中的抽象推演. 若无法明确热点所在位置, 则难以将冷却结构布置在最需要的位置. 这也正是DARPA于2023年启动Thermonat项目的核心动因——该项目旨在从单个CMOS晶体管入手, 构建跨尺度的电-热耦合模型.

这里需要特别指出, 在金属和重掺杂半导体中电子是不可忽视的热载体: 芯片的金属互连线中, 自由电子对热量传递的贡献往往远超声子导热; 在高功率器件的源漏接触区等局部重掺杂区域, 高浓度自由电子与声子发生强烈散射, 既显著降低了晶格热导率, 又通过维德曼-弗兰兹(Wiedemann-Franz)定律贡献了额外的电子热导, 电子-声子耦合效应显著影响局部热点的形成与扩散. 而在芯片所用的绝大部分半导体材料(如硅)中, 由于载流子浓度相对较低, 声子承担了主要的热输运任务——从晶体管沟道的热扩散, 到封装层的热耗散, 声子的产生、传播与散射决定了芯片的温度分布——是芯片热管理的核心作用对象.

3.2 界面: 热的"关卡"

若将声子输运比作热的"长途旅行", 则界面可视为热的"关隘卡口". 界面热阻, 亦称卡皮查热阻(Kapitza resistance), 由苏联物理学家彼得·卡皮查于1941年首次发现并命名. 他在测量液氦与固体铜之间的传热过程时观察到: 热量流经界面时, 会在界面两侧形成显著的温度跳变, 如同界面本身对热流施加了额外阻力[7], 具体参见图4. 界面热阻的数学定义较为简洁, 即

$$R = \Delta T / q, \tag{1}$$

其中, ΔT为界面温差, q为通过该界面的单位面积热流. 然而, 其物理机制却极为复杂. 当声子从材料A传播至材料B时, 会经历以下三种情况: (1) 声学失配. 两种材料的声子频谱、声

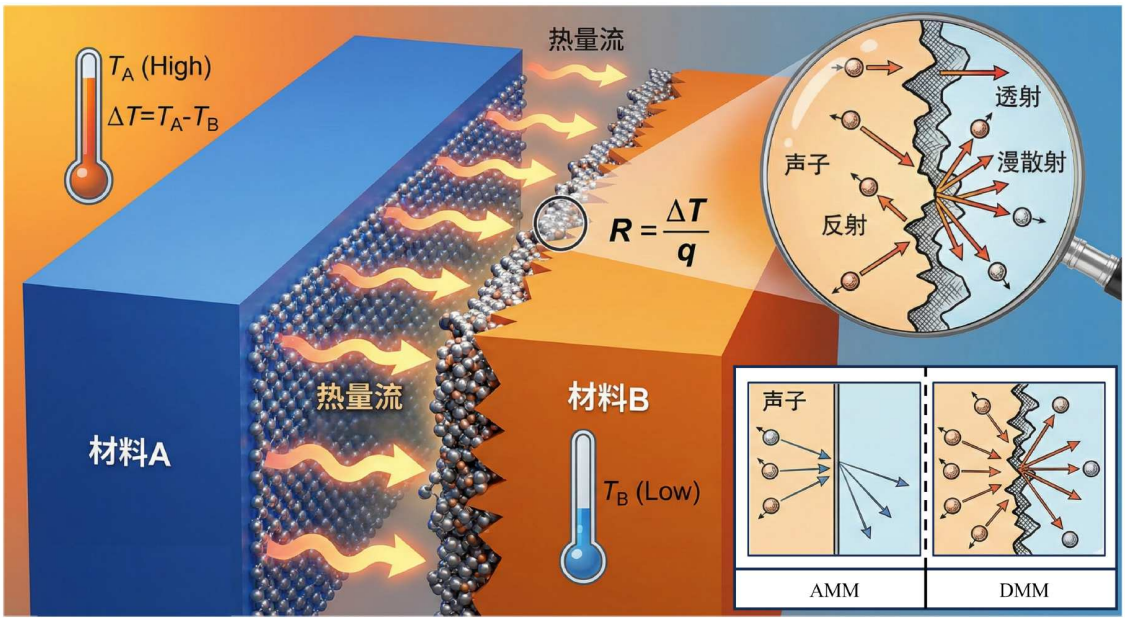


图 4 界面热阻——热量传递的“关卡”. 当热量流经两种材料的界面时, 声子会发生反射、透射及漫散射现象, 致使界面两侧出现温度跳变ΔT, 据此定义界面热阻$R=\Delta T/q$. 右下图展示了声学失配模型(AMM)与漫散射失配模型(DMM)用于预测界面热阻的物理图像. 其中, AMM假定界面呈原子级平整状态, 声子的传播遵循斯涅耳(Snell)定律; DMM则假定声子会发生完全漫散射. 由于真实界面存在原子互混、化学键合及局部应力等情况, 导致实测值与理论值之间存在30%~50%的偏差. 本图基于作者原创思路, 采用AI工具绘制

Figure 4 Interface thermal resistance — the “barrier” of heat transfer. When heat flows across the interface between two materials, phonons undergo reflection, transmission and diffuse scattering, resulting in a temperature jump ΔT across the interface. Accordingly, the interface thermal resistance is defined as $R = \Delta T/q$. The bottom-right figure shows the physical pictures of the acoustic mismatch model (AMM) and diffuse mismatch model (DMM) for predicting interface thermal resistance. Specifically, AMM assumes that the interface is atomically flat, and phonon propagation follows Snell’s law; DMM assumes that phonons undergo completely diffuse scattering. Owing to atomic intermixing, chemical bonding and local stress at the real interface, there exists a deviation of 30%–50% between the measured values and theoretical predictions. This figure is based on the author’s original idea and is drawn using AI tools

速及密度存在差异. 如同光从空气射入水中时会在界面发生折射与反射, 声子在不同材料的界面处同样会出现部分反射现象. (2) 漫散射. 实际界面并非原子级平整, 存在原子台阶、缺陷、氧化层及非晶层等结构. 当声子入射至此类粗糙结构时, 会产生随机漫散射. (3) 非弹性过程. 声子可在界面处发生“分裂”或“合并”, 即单个高频声子转化为两个低频声子, 或两个低频声子合并为一个高频声子. 这一过程涉及能量与动量的重新分配.

经典理论模型AMM与DMM, 对界面热阻的预测结果往往与实验值存在30%~50%的偏差[5]. AMM假设界面为原子级平整的“镜面”, 声子按斯涅耳定律折射透射; DMM则假设声子遭遇界面后完全丧失原有相位信息, 向各个方向随机散射. 然而, 真实界面远非如此理想: 芯片制造过程中的沉积、键合与退火会在界面处形成纳米级厚度的原子互混层或非晶过渡区, 该区域声速与密度呈连续渐变而非突变; 同时, 界面处的化学键重组、氧化层残留及局部应力场会引入额外的声子散射中心. 这些因素共同导致声子在宽频范围内的透射效率显著低于理想模型假设, 实测界面热阻也往往偏离AMM与DMM的理论预测值. 这对于集成电路散热而言, 在从芯片到环境所构建的热流路径上, 每一个“两种材料接触”之处均构成一个热的“关卡”. 具体而言, 晶体管/硅衬底界面、硅衬底/热界面材料、热界面材料/芯片封装盖、封装盖/外部热界面材料以及外部热界面材料/散热器底板均为“关卡”[9]. 这些“关卡”的传热效率直接决定了芯片的温度. 并且, 每个“关卡”的物理机制尚未得到全面且深入的理解. 这正是热管理作为“基础学科”的根基所在: 它并非一个发展成熟、问题已全部解决的领域, 而是蕴含着诸多尚未解答的根本性问题.

3.3 尺度: 从纳米到千公里, 十余个数量级的统一战场

热管理工程师的思维需在十余个数量级范围内进行跨尺度切换[18], 如图5所示: 从纳米声子输运到千公里卫星热控, 五个典型尺度对应密度泛函理论、两相流、热阻网络、CFD与辐射换热五种截然不同的方法论. 纳米尺度: 需分析声子在硅-锗界面的透射系数, 这依赖密度泛函理论计算, 并求解含时间变化的玻尔兹曼方程[5,20,21]. 微米尺度: 需考量芯片背面蚀刻微通道内的流动与相变过程, 这依赖两相流动力学分析, 且需运用临界热流密度模型[24,25]. 毫米尺度: 需考量

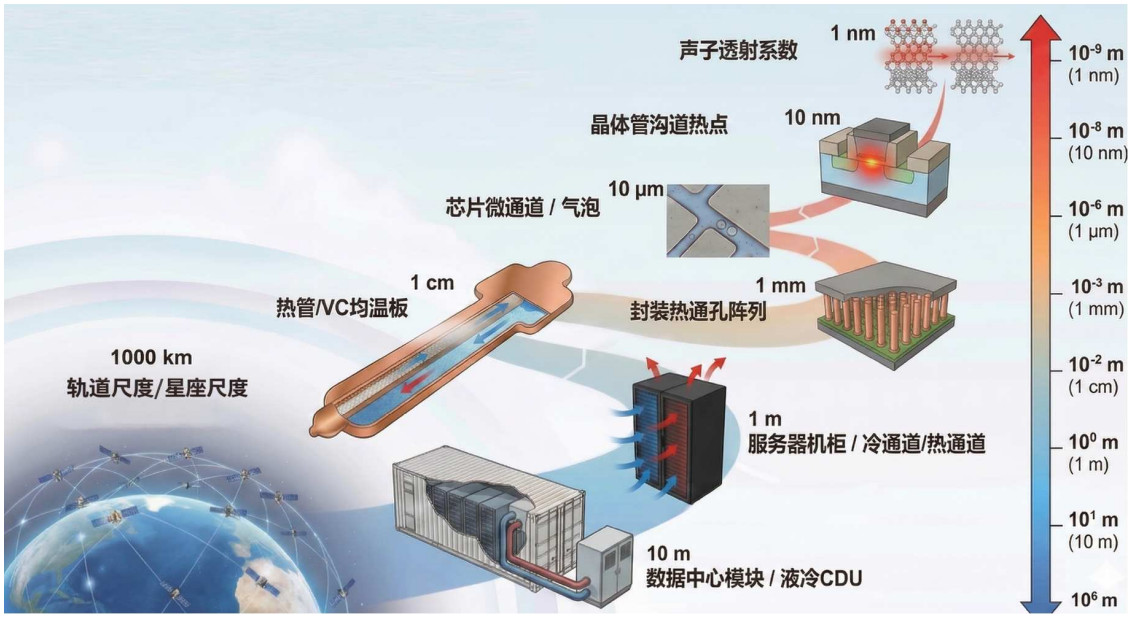


图 5 热管理的十余个数量级尺度全景. 从1 nm至1000 km, 热管理工程师的思维需在十余个数量级间切换. 纳米尺度: 声子透射系数(密度泛函理论); 微米尺度: 芯片微通道沸腾(两相流); 毫米尺度: 封装热通孔阵列(热阻网络); 米尺度: 数据中心机柜气流组织(CFD); 1000 km尺度: 轨道尺度/星座尺度(此处1000 km指卫星星座轨道高度及空间调度范围, 而非散热对象本身的物理尺寸). 热管理于不同尺度呈现全然不同的问题特征. 本图基于作者原创思路, 采用AI工具绘制

Figure 5 Panorama of thermal management across more than ten orders of magnitude. From 1 nanometer to 1000 kilometers, thermal management engineers need to switch thinking across more than ten orders of magnitude. Nanoscale: phonon transmission coefficient (density functional theory); Micrometer scale: microchannel boiling of chips (two-phase flow); Millimeter scale: thermal via array in packaging (thermal resistance network); Meter scale: airflow organization in data center cabinets (CFD); 1000 km scale: orbital/constellation scale (here, ~1000 km refers to the orbital altitude and spatial coverage of the satellite constellation, not the physical size of the thermal management object itself). Thermal management exhibits completely different problem characteristics at different scales. This figure is based on the author's original idea and is drawn using AI tools

封装基板内铜柱热通孔的布局设计, 这依赖热阻网络分析, 且需开展多热源耦合优化. 米尺度: 需考量数据中心机柜的气流组织形式, 这依赖计算流体力学方法, 且需采用冷热通道隔离设计[12]. 千公里尺度: 需考量低轨卫星星座的热控调度, 这依赖轨道热环境建模, 且需进行辐射换热网络求解. 在这里, 无法通过对流散热, 也无地板可供架空, 热量只能通过辐射缓慢消失, 而太阳辐照每90分钟引起一次剧烈的温度循环.

同一研究对象与目标, 在十余个数量级的尺度下呈现出全然不同的特征[3]. 很少有学科要求研究者如此切换认知视角. 这正是热管理领域对思维能力的极致锤炼: 研究者既要能深入探究声子输运的微观机制, 又要能宏观把握整个数据中心的能量流动规律.

4 集成电路: 热管理的主战场

若将热管理视为一门学科, 那么集成电路堪称其主战场、试验场与发源地. 在过去三十年里, 热管理领域的每一项重大突破, 几乎均由集成电路散热需求所驱动; 而当下, 热管理已开始反向重塑集成电路的设计方法论.

4.1 功耗墙: 摩尔定律的终结者?

约2005年, 半导体产业经历了一个隐蔽却致命的转折. 登纳德缩放定律(Dennard scaling)失效[16](详见图6). 登纳德缩放定律作为摩尔定律的"经济基石", 其内容为: 当晶体管尺寸缩减至原有尺寸的70%(面积缩小一半)时, 工作电压亦同比降低30%, 此时单位面积的功耗将维持恒定. 这表明能够无限制地缩小晶体管尺寸、提升集成密度, 而无需担忧芯片因过热而损坏. 不过, 登纳德缩放定律存在一个潜在假设, 即晶体管的阈值电压可随工作电压同比降低. 然而, 阈值电压无法无限制地降低, 当阈值电压低于0.2~0.3 V时(工作电压通常需维持为阈值电压的2.5~4倍以保证电路性能, 如图6所示), 晶体管在"关断"状态下无法实现完全关断, 漏电流呈指数级增长.

2005年左右, 业界达到了这一临界点. 此后, 晶体管尺寸持续缩小, 但工作电压基本不再下降, 其结果为: 单位面积的功耗开始随晶体管密度同步急剧上升(为应对热管理瓶颈, 产业被迫从单核频率扩展转向多核并行架构, 通过降低单核工作电压和运行频率, 使功率密度在2005年后有所回落, 进入"双核时代"). 这便是"功耗墙"(Power Wall)的物理根源[15].

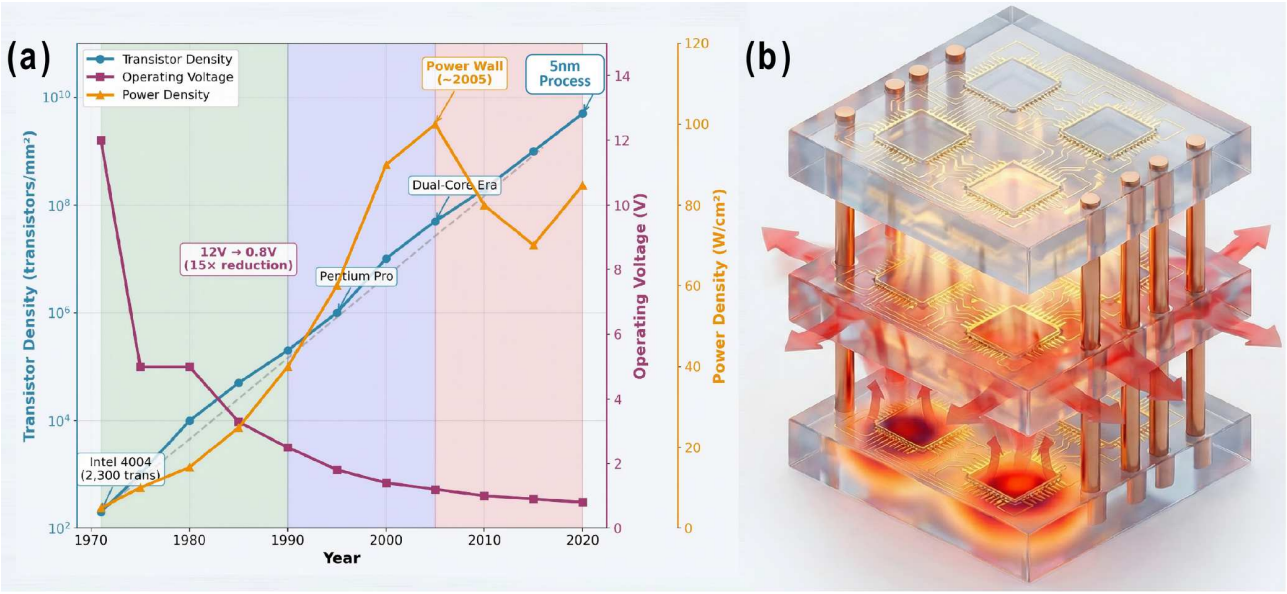


图 6 功耗墙与三维集成的热阱困境. (a) 功耗密度的历史发展趋势. 1974~2005年, 登纳德缩放效应有效, 功耗密度得以控制, 但仍随晶体管性能提升而逐步增长; 自2005年左右登纳德缩放失效、工作电压无法继续降低后, 功耗密度攀升至"功耗墙"峰值, 产业被迫转向多核架构, 通过降低单核电压和频率使整体功耗密度有所回落. 此后, 随着集成度持续提升, 芯片功耗密度再次攀升, 至2020年代已接近100 W/cm^2, 而先进AI芯片的局部热点区域功耗密度甚至可达300 W/cm^{2}以上. (b) 三维堆叠芯片热阱效应的示意图. 下层芯片被上层芯片所覆盖, 热流仅能通过边缘的狭窄路径排出, 散热难度随堆叠层数呈非线性增长. 本图基于作者原创思路, 采用AI工具绘制

Figure 6 The power wall and thermal trap dilemma of 3D integration. (a) Historical development trend of power density. From 1974 to 2005, benefiting from Dennard scaling, power density was kept under control but still increased gradually with transistor performance improvement. After Dennard scaling broke down around 2005 and operating voltage could no longer be reduced, power density climbed to the peak of the "Power Wall", forcing the industry to shift toward multi-core architectures. By reducing per-core voltage and frequency, the overall power density declined somewhat. Subsequently, as integration density continued to increase, chip power density rose again, approaching nearly 100 W/cm^2 by the 2020s, while local hotspot power densities in advanced AI chips can even exceed 300 W/cm^2. (b) Schematic of the thermal trap effect in 3D stacked chips. The lower chip is covered by the upper chip, and heat can only be dissipated through narrow pathways at the edges, with cooling difficulty increasing non-linearly with the number of stacked layers. This figure is based on the author's original idea and is drawn using AI tools

2024年, 最先进的人工智能加速芯片——英伟达Blackwell架构图形处理器(graphics processing unit, GPU), 单颗总功率已达1200 W; 对于其下一代Rubin架构, 行业普遍预测其总功率将突破2000 W[14].

2000 W功率的物理意义可作如下类比. 通常, 家用电磁炉的最高功率档位为2200 W. 若以电磁炉功率对指甲盖大小的芯片持续加热, 其局部热流密度将超过300 W/cm^2, 此数值为数倍于典型核反应堆堆芯的热流密度. 传统风冷技术的极限散热热流密度约为100 W/cm^2 [12]. 这表明, 单纯依赖外部散热技术的升级, 已难以应对芯片自身功耗密度的增长.

4.2 三维集成: 把热阱变成热路

在功耗密度急剧攀升的同时, 芯片架构设计人员还采取了一项加剧这一状况的技术路径: 三维集成[26]. 逻辑芯片从平面架构向三维架构转变具有合理性: 相较于将多个芯片并列铺设于载板之上, 并采用较长的互连线进行信号传输, 将芯片垂直堆叠, 并利用硅通孔(through-silicon via, TSV)实现直接连接更为高效. 通过这种方式, 互连长度能够从毫米级别缩减至微米级别, 从而同步降低信号延迟与互连功耗. 然而, 三维集成由此引发了严峻的热管理挑战: 三维堆叠造成多层异质界面的累积热阻呈指数级增长, 当计算密度提升至原来的三倍时, 散热难度则增加至原来的十倍. 工程师们从两个方向寻求解决方案, 其核心思路均聚焦于降低界面累积热阻、优化声子输运路径.

方向一: 在堆叠内部预设热通道. 在芯片设计阶段, 将高导热材料"嵌入"三维堆叠结构. 斯坦福大学研究团队提出三维热支架(3D thermal scaffold): 在晶体管层下方嵌入纳米厚度的金刚石框架, 热量从热点直接横向扩散至金刚石支架, 再向下传导至衬底或向上传导至封装盖, 减少多层堆叠产生的界面数量, 从根源上降低累积热阻[26].

方向二: 将冷却结构集成到封装. 此为"近结点冷却"(near-junction cooling)的核心思想: 并非将热量导出芯片后再处理, 而是在芯片背面几十微米处直接蚀刻微通道, 使冷却液贴近发热区流动[10], 缩短热流路径, 减少热流经过的界面数量, 进而降低累积界面热阻对散热的限制. 瑞士洛桑联邦理工学院的研究表明, 在三维集成电路(three-dimen-

sional integrated circuit, 3D-IC)和量子计算等对温度极为敏感的场景中，硅微通道冷板可实现显著优于传统液冷的散热能力, 热流密度处理上限被提升至500 W/cm^{2}以上[10].

4.3 材料革命: 界面、热沉与“二合一”半导体

当传统铜铝接近传热极限，新材料正在从实验室走向晶圆厂. 这场材料革命沿着三条路径展开: 高性能界面材料、高性能热沉材料，以及一个更激进的方向——让半导体材料自身成为极限导热载体. 这三条路径, 见图7, 均源于对声子输运特性的精准调控，分别对应着热管理材料的三个终极追求: 界面输运、热流扩散、源端治理. 它们都在从实验室走向工程化的路上，而这条路，正是新质生产力最活跃的试验场.

(1) 高性能界面材料: 打通热的“咽喉”. 芯片与散热器间的界面是热流路径上最窄的“咽喉”，接触热阻在此处通常占据总温降的一半以上[9]. 理想的界面材料需具备极薄、极柔且高导热的特性, 石墨烯与碳纳米管是该领域的有力竞争者. 单层石墨烯的本征热导率可达3000~5000 W/(m K), 为铜的10倍以上; 碳纳米管的理论热导率同样十分可观[27]. 但核心挑战在于如何将纳米尺度的优异性能转化为宏观可用的界面层[28]. 研究者正从两条技术路径寻求突破: 其一, 将石墨烯作为填料添加至金属或聚合物基体中，制备导热复合界面材料[29]; 其二，利用垂直排列的碳纳米管阵列直接构建散热通道, 借助其轴向超高热导特性打通界面[28]. 不过, 碳基材料仍面临若干难题，如与基底间的界面热阻、定向排列的工艺控制及量产成本等，每一项问题的解决均需围绕声子输运与界面散射的核心物理机制展开优化.

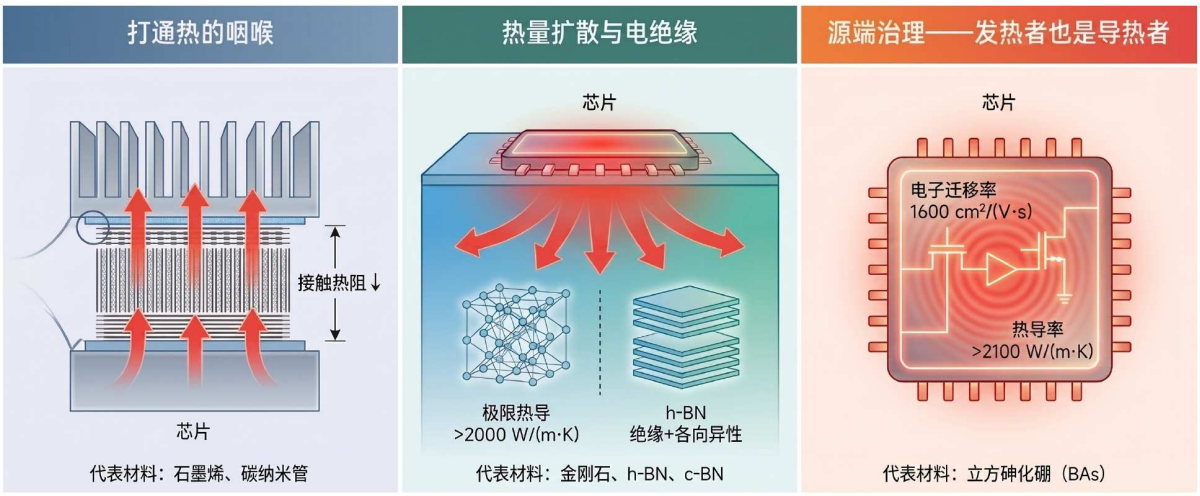
(2) 高性能热沉材料: 金刚石与氮化硼家族. 当热量从芯片传出时，需实现快速扩散，这是热沉材料的核心功能——其不一定直接接触芯片，却对热量能否快速疏散起着决定性作用. 金刚石在该领域处于领先地位，其热导率高达2200 W/(m K)以上(碳原子通过强sp^{3}共价键构成三维网络, 使声子群速度极高, 并显著抑制声子-声子散射), 为铜的5倍, 且热膨胀系数与半导体材料相匹配. 人造金刚石薄膜已应用于高功率激光器、射频器件等领域. 然而，金刚石亦存在自身局限: 成本高昂、大尺寸单晶生长难度大、加工工艺复杂[30]. 另外，还需要优化降低金刚石与芯片衬底之间的界面热阻, 以避免界面导热成为新的瓶颈.

在诸多应用场景中, 热沉材料还需满足另一项严苛要求, 即电绝缘性. 芯片背面与散热结构直接接触，任何导电通路均可能引发短路失效. 六方氮化硼(h-BN)素有“白石墨”之称, 其结构与石墨烯相似，却是性能优异的电绝缘体. 与金刚石结构类似, 六方氮化硼依靠B-N共价键实现高声速, 同时因层状结构呈现出极强的各向异性——面内B-N共价键网络支撑高效声子传输，而层间弱范德华力则导致声子严重局域化; 其面内热导率可达约300 W/(m K)，面外热导率则相对较低. 这一特性既是挑战, 也是机遇: 通过取向调控, 可使热量沿特定方向输运，同时维持垂直方向的电绝缘[31]. 更值得关注的是立方氮化硼(c-BN), 它具有类金刚石的闪锌矿结构, 热导率仅次于金刚石，同时是禁带宽度最大的半导体材料之一. 这意味着，c-BN不仅可作为热沉材料，更是下一代功率器件的重要候选材料，它有望同时承担“发热体”与“散热体”的双重角色. 金刚石与氮化硼家族构成了热沉材料的两端: 一端追求极限热导，一端追求导热与绝缘的统一. 而中间地带的铜/金刚石复合材料、铝/碳化硅复合材料等, 则在成本与性能之间寻求平衡.

(3) 源端治理: 砷化硼——热管理与电性能二合一的半导体. 若前两条路径旨在“优化热路径”，则第三条路径更具根本性: 让发热的源头自身成为高效导热体[32,33]. 立方砷化硼(BAs)是近年来最引人瞩目的发现之一. 2013年，理论预测其热导率可达2000 W/(m K)以上，与金刚石相当. 2025年10月, 休斯顿大学团队报道: 通过优化合成工艺, 高质量砷化硼单晶的室温热导率实测值已超过2100 W/(m K)，甚至高于金刚石，成为各向同性材料中的导热新纪录. 砷化硼的超高热导率源于其独特的声子谱特性: 光学支与声学支之间的大频率间隙抑制了三声子散射，使声子具有极长的平均自由程. 更关键的是砷化硼的“双重属性”: 其电子与空穴迁移率均高达1600 cm^2/(V s), 远超硅; 禁带宽度适中, 热膨胀系数与常见半导体匹配. 这意味着: 既可将其用于晶体管制造，又能通过自身实现热量导出，这是硅、碳化硅、氮化镓均无法实现的特性[32,33]. 砷化硼距离产业化仍有差距: 高质量单晶生长难度大, 缺陷浓度较高, 掺杂工艺尚未成熟. 但它已明确指出一个方向: 热管理与电性能的协同设计，可在材料本源层面实现内禀耦合.

4.4 方法论革命: 从“后端散热”到“前端热设计”

上述诸多变化相互叠加，正催生一场更为深刻的范式转移. 传统集成电路设计流程呈线性特征，具体为: 架构设计—逻辑设计—电路设计—物理设计—流片—封装—散热方案选型. 在此流程中, 热管理始终处于最后环节, 被动承接前述所有环节产生的热量，并设法将其耗散. 该流程已不再适用[15]. 2025年设计自动化大会(Design Automation Conference, DAC)上, 多位专家明确指出: 冷却策略必须前置到设计阶段，从器件布局、布线、电源分配网络布局等层面开始优化[14]. 具体包含: (1) 温度感知设计: 电子设计自动化(electronic design automation, EDA)工具开始内建温度感知模型, 自动识别高功耗模块，避免在物理布局阶段将它们集中布置. (2) 电源分配网络优化: 通过动态电流分布建模, 预测并消除局部电流过载引发的热点. (3) 芯粒异构集成: 将高热元件分散在硅中介层上，通过热通孔阵列为每一颗芯粒构建专属冷却路径[16]. (4) 热-功协同调度: 芯片运行时, 操作系统根据片上温度传感器阵列反馈，动态调度计算任务、调整核心电压



路径	核心价值	解决的关键问题
界面材料	打通界面咽喉	接触热阻、界面输运效率
热沉材料	热量扩散与电绝缘	热扩散能力、电隔离需求
二合一半导体	源端治理	从发热源头解决热问题

图 7 热管理材料革命的三条路径. 热管理新材料的探索沿三条并行路径推进. 左图: 高性能界面材料(如石墨烯、碳纳米管)旨在打通芯片与散热器之间的“咽喉”, 降低接触热阻. 中图: 高性能热沉材料中, 金刚石追求极限热导率, 氮化硼家族(h-BN、c-BN)则在具备导热性能的同时实现电绝缘. 右图: 立方砷化硼(BAs)代表更为激进的“二合一”方向, 其既是高热导材料(热导率>2100 W/(m K)), 又是高电子迁移率半导体, 使发热源头自身成为导热优势载体. 三条路径分别对应界面输运、热流扩散、源端治理三大目标, 共同构成热管理材料革命的完整图景. 本图基于作者原创思路, 采用AI工具绘制

Figure 7 Three parallel paths in the material revolution for thermal management. The exploration of new materials for thermal management proceeds along three parallel paths. Left: High-performance thermal interface materials (e.g., graphene, carbon nanotubes) aim to unblock the “bottleneck” between the chip and the heat spreader, reducing contact thermal resistance. Middle: Among high-performance heat sink materials, diamond pursues ultimate thermal conductivity, while the boron nitride family (h-BN, c-BN) provides thermal conduction with electrical insulation. Right: Cubic boron arsenide (c-BAs) represents a more radical “two-in-one” concept—it is both a high thermal conductivity material (>2100 W/(m K)) and a semiconductor with high electron mobility, turning the heat source itself into a high-performance heat conduction medium. These three paths correspond to the three major objectives of interfacial transport, heat flux spreading, and source-side governance, respectively, jointly forming a complete landscape of the material revolution in thermal management. This figure is based on the author’s original idea and is drawn using AI tools

频率[34]. (5) 散热能力就是算力: 在功耗墙时代, 芯片的峰值性能不再由晶体管数量决定, 而由热管理系统能维持的稳定功耗水平决定. 散热能力直接转化为算力. 这是一个根本性的范式转移.

5 AI数据中心: 热管理成为算力基础设施

如果说集成电路是热管理的“芯片级”战场, 那么AI数据中心就是热管理的“系统级”战场. 这两个战场的尺度相差六个数量级, 但它们的底层物理逻辑是贯通的.

5.1 功率密度跃迁: 每机架10~250 kW

走进一座传统数据中心, 首先映入眼帘的往往是噪音——成千上万个风扇发出的低沉嗡鸣声: 芯片旁的风扇、机柜背板的风扇、交换机上的风扇. 这些风扇不断将芯片产生的热空气输送至空调系统, 以维持计算芯片运行. 这套体系正在系统性失效[12]. 过去五年, 最先进AI芯片的功耗呈急剧增长[14], 如表2所示.

单颗芯片功耗的跃迁, 叠加单机柜部署密度的影响, 导致机柜功率密度呈现数量级跃升[12]: 传统云计算数据中心的设计基准为单机架功率密度5~10 kW/rack; 当前, 英伟达GB200 NVL72系统机架功率密度达140 kW/rack, 微软Azure最新部署的单机架功率密度达50 kW/rack, 谷歌的张量处理单元(tensor processing unit, TPU)集群正推行60 kW/rack的机架密度标准; 对于面向未来的AI训练集群, 行业已普遍将250 kW/rack确立为下一代设计目标[10], 详见图8.

数字的对比结果表明: 传统架空地板空气冷却系统在机架功率超过25 kW/rack时将完全失效[12]. 热空气回流可能引发“热失控”现象——温度持续上升, 最终导致设备强制降频甚至停机. 一个40 kW的机架所产生的热量, 相当于14台家用取暖器持续满负荷运行时的发热量. 若将8个这类机架并排

表 2 先进AI芯片的总功率

Table 2 Total power of advanced AI chips

年份	型号	总功率(W)
2017年	英伟达V100 GPU	300
2020年	英伟达A100 GPU	400
2022年	英伟达H100 GPU	700
2024年	英伟达Blackwell GPU	1200
2025年	英伟达B300 GPU	1400
未来	下一代Rubin	>2000

布置, 需在不足20 m^2的空间内处理相当于一栋小型办公楼的热量输出.

5.2 液冷的物理必然性

当空气冷却受限于物理极限时, 液冷技术成为高热流密度场景下的主流可行路径[10]. 从数学原理来看, 液冷技术具备显著优势[3]: 水的比热容为空气的4倍, 密度为空气的800倍, 因此单位体积水的吸热能力可达空气的3200倍; 同时, 水的导热系数为空气的23.5倍. 将手伸入热烤箱通常不会造成灼伤, 但伸入沸水则会立即导致严重烫伤, 这是因为液体的传热速度远快于空气. 此外, 液冷技术凭借极高的热容密度, 仅需较小流量即可输送大量热量, 驱动循环所需的泵功远低于风冷系统为输送巨量空气而消耗的风扇或风机电能, 液冷系统的能源效率优势同样突出. 电能利用效率(power usage effectiveness, PUE)是衡量数据中心系统级能效的核心指标. PUE定义为数据中心总用电功率与IT设备用电功率之比, 其数值越接近1, 表明冷却、供配电等非IT能耗占比越低. 传统风冷数据中心PUE通常在1.5~1.8之间, 即IT设备每消耗1 W计算用电, 需额外消耗0.5~0.8 W的电用于冷却[12]; 而先进液冷系统的PUE可降至1.02~1.10[10]. 对于一座10 MW的数据中心而言, PUE每优化0.1, 每年可节约电费约680万元人民币(约合100万美元).

5.3 液冷技术的四条技术路线

液冷并非单一技术, 而是一个技术谱系. 依据冷却液是否与芯片直接接触、是否发生相变, 当前数据中心液冷技术可划分为四条主要技术路线[10](图9). 所有技术路线的设计核心, 均围绕在系统层面减少热流路径界面热阻、使芯片级声子输运产生的热量更高效地传递至冷却介质这一目标展开. 图9以星级符号直观比较各路线的技术成熟度, 星级越高代表该路线在工程部署、标准制定及供应链配套方面越完善.

(1) 路线一: 单相直触芯片液冷. 此路线为目前技术成熟

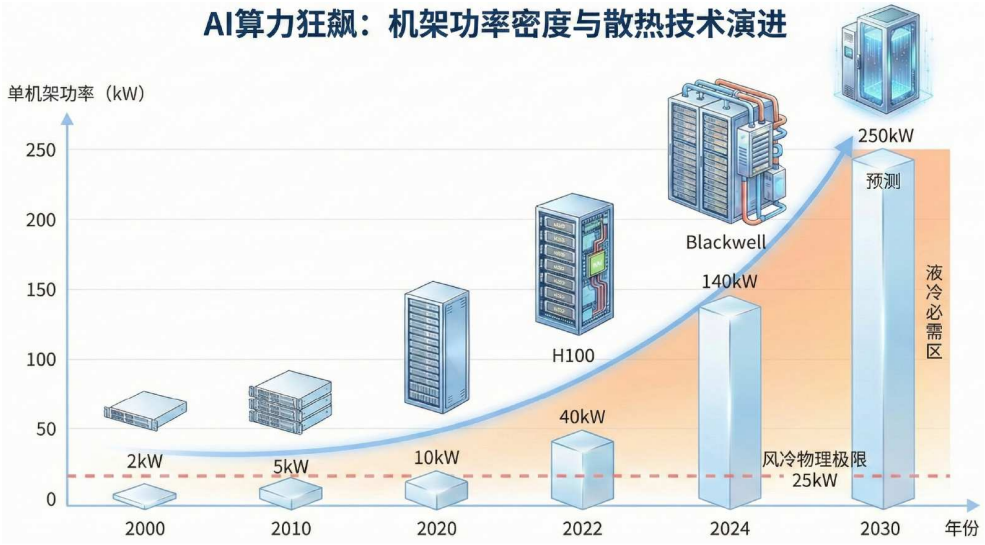


图 8 数据中心功率密度跃迁(2000~2030年). 2000~2020年间, 数据中心机架功率密度由2 kW/rack逐步提升至10 kW/rack, 传统风冷技术足以满足散热需求. 2020~2025年, 人工智能训练集群规模快速扩张, H100集群功率密度突破40 kW/rack, Blackwell集群达140 kW/rack. 2030年预期目标为250 kW/rack. 风冷技术的物理极限约为25 kW/rack, 在高热流密度场景下, 液冷遂成为主流技术路径. 本图数据来源于IDTechEx《数据中心热管理2026~2036》报告, 采用AI工具绘制

Figure 8 Power density evolution in data centers (2000–2030). From 2000 to 2020, the rack power density in data centers increased gradually from 2 kW/rack to 10 kW/rack, and traditional air cooling was sufficient to meet the cooling demand. From 2020 to 2025, driven by the rapid expansion of artificial intelligence training clusters, the power density of H100 clusters exceeded 40 kW/rack, and that of Blackwell clusters reached 140 kW/rack. The target for 2030 is expected to be 250 kW/rack. The physical limit of air cooling is approximately 25 kW/rack, and liquid cooling has become the mainstream technical solution in high heat flux scenarios. This figure is based on data from the IDTechEx report "Thermal Management for Data Centers 2026–2036" and is drawn using AI tools

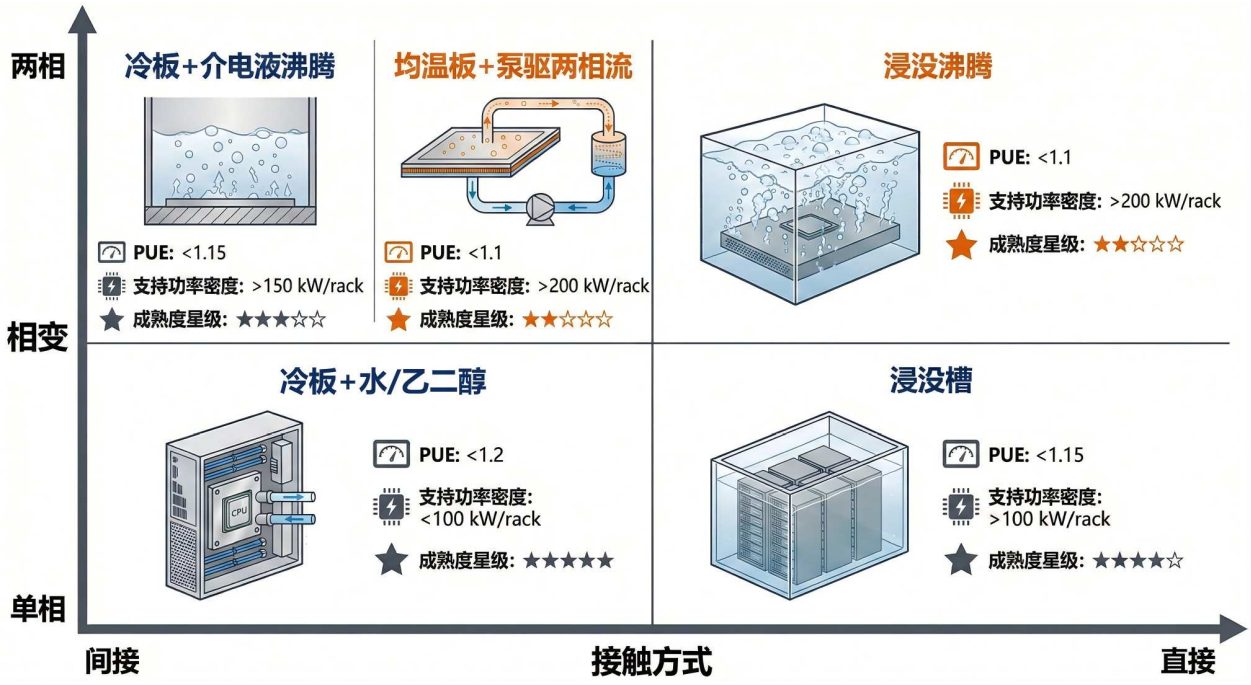


图 9 液冷技术路线谱系. 依据冷却液是否与芯片直接接触、是否发生相变, 液冷技术可划分为四条路线. 单相直触路线(★★★★★)采用冷板与水乙二醇组合, 成熟度处于最高水平; 两相直触路线(★★★☆☆)采用冷板与介电液沸腾组合, 散热系数得以提升; 单相浸没路线(★★★★☆)将整机浸入油性介质, 自然对流效果有限; 两相浸没路线(★★☆☆☆)将整机浸入低沸点介电液, 借助沸腾相变, 可支持250 kW/rack的散热需求. 本图基于作者原创思路, 采用AI工具绘制

Figure 9 The spectrum of liquid cooling technologies. According to whether the coolant directly contacts the chip and whether a phase change occurs, liquid cooling technologies are classified into four routes. Single-phase direct contact (★★★★★): cold plate with water-glycol mixture, with the highest maturity; two-phase direct contact (★★★☆☆): cold plate with dielectric fluid boiling, with improved heat dissipation coefficient; single-phase immersion (★★★★☆): the entire server immersed in oil-based medium, with limited natural convection; two-phase immersion (★★☆☆☆): the entire server immersed in low-boiling-point dielectric fluid, supporting heat dissipation of 250 kW/rack via boiling phase change. This figure is based on the author's original idea and is drawn using AI tools

度最高、部署最广泛的方案. 将一块金属冷板直接安装在芯片表面, 冷板内部集成微通道, 冷却介质(通常为水与乙二醇的混合液)流经微通道以带走热量. 该方案的现实局限在于: 无法为服务器内所有发热部件安装冷板. 实践中仅对GPU、中央处理器(central processing unit, CPU)等高功耗芯片采用直触液冷, 而电源模块、内存等中低功耗部件仍沿用风冷方式. 这是一种"混合冷却"架构, 其中液冷承担约80%的散热负载, 风冷承担剩余20%的散热负载.

(2) 路线二: 两相直触芯片液冷. 当单相水冷技术趋近传热极限时, 物理学为其提供了第二种技术路径——相变潜热的利用[24]. 该技术方案采用低沸点介电型冷却液(其沸点可低至18℃), 使冷却液在冷板内部、芯片表面直接发生沸腾, 借助汽化潜热实现大量热量的转移, 且在此过程中温度保持恒定. 由于冷却液仅发生相态变化而温度不升高, 机房冷却系统可接受的回水温度较单相系统高出6~8℃, 进而显著降低能耗[10]. 此外, 两相系统的流量需求仅为单相系统的五分之一, 不仅泵送能耗更低, 对设备产生的机械应力也更小.

(3) 路线三: 单相浸没液冷. 绕开冷板, 将整台服务器直接"浸没"在盛有介电冷却液的槽体中. 服务器内部所有发热部件——芯片、内存、电源、硬盘——均由同一种液体直接冷却. 单相浸没面临的挑战在于: 油基冷却液的自然对流换热系数有限. 当先进芯片的功率密度超出一定阈值时, 必须为这类芯片额外配置冷板以强化局部流动换热, 这意味着系统需同时维护两套冷却回路[10].

(4) 路线四: 两相浸没液冷. 这是当前数据中心液冷领域最具突破性的技术路线[12]. 整台服务器浸入低沸点介电液中, 高温部件表面直接沸腾, 借助相变潜热实现等温冷却. 从液体到蒸汽的相变过程可吸收大量热量, 理论上可支持每机架250 kW以上的功率密度[10].

当单一液冷技术的性能趋近极限时, 组合方案成为新的技术突破口. 组合方案以高导热绝缘热沉或高功率均温板作为"热扩展中继层", 与冷板或浸没冷却技术协同运行. 在"均温板+冷板"方案中, 均温板先将芯片热点的热量快速横向扩展, 使热流密度均匀化, 再由冷板内的单相或两相流将热量带走, 这相当于为芯片增设了热扩散增强层. 在"均温板+浸没"方案中, 均温板可将芯片热量高效传导至浸没液体中, 有

效解决浸没冷却技术中高功率芯片局部换热不足的问题. 相关研究显示, 此类组合增强方案能够使散热能力提升30%以上, 同时降低泵送功耗[10]. 单相直触方案成熟可靠, 核心是减少芯片与冷却介质的界面热阻; 两相直触效率跃升, 核心是在低界面热阻基础上利用相变潜热; 单相浸没覆盖全面, 核心是消除人工界面、实现热流直接传递; 两相浸没潜力巨大, 核心是界面热阻最小化与相变高效吸热的结合; 组合增强则融合众长、突破极限, 核心是通过热扩展中继层优化传热路径、进一步降低界面热阻. 这几条路线并非替代关系, 而是面向不同单机架功率密度和应用场景的技术谱系——从100 kW/rack到>250 kW/rack, 热管理工程师始终拥有多种技术选择.

5.4 冷却经济学: 从资本支出到运营支出的财务决策

极高密度冷却不仅是技术层面的决策, 更是财务层面的决策[12]. 从资本支出(capital expenditure, CapEx)角度看, 传统风冷具有一定优势; 但从运营支出(operating expenditure, OpEx)维度分析, 情况则有所不同[10]. 液冷系统的功耗较风冷系统降低30%~40%, 以1 MW的部署规模为例, 仅电费一项每年即可节省270万至340万元人民币(约合40万至50万美元). 液冷系统可维持更低温度, 减少热循环应力, 使设备寿命延长20%~30%, 从而降低设备更换成本; 同时, 液冷技术支持更高的部署密度, 能够在现有设施中实现更多计算量, 避免标准数据中心每兆瓦约6820万至1亿元人民币(约合1000万至1500万美元)的新建成本.

更为隐性的成本体现为机会成本. 以OpenAI的生成式预训练变换器(generative pre-trained transformer, GPT)模型训练运行为例, 若缺乏经过优化的冷却基础设施, 其耗时将达到现有水平的10倍[14]. 每机架40 kW与100 kW的散热能力差异, 直接决定了下一代大模型的训练周期为几周或几个月. 散热能力可直接转化为研发效率, 进而转化为市场先发优势, 这正是热管理作为“新质生产力”的核心体现.

5.5 下一个前沿: 芯片与数据中心的热协同

芯片与数据中心热协同(thermal co-design of chips and data center)是指打破传统“芯片级热管理”与“数据中心级热管理”之间的物理与信息壁垒, 通过统一的热学建模、实时数据互通与跨尺度优化算法, 实现从晶体管结温到冷却塔散热的全链路热流统筹调控.

2025年, 一场深刻变革正在发生: 芯片级热管理与数据中心级热管理的边界正逐步消失[10,34], 而二者融合的核心, 在于依托声子输运与界面热阻两大基础物理概念, 构建从芯片延伸至数据中心的全链路热流优化体系. 从历史实践来看, 这两个层面曾处于解耦状态: 芯片设计团队仅关注从结温至封装盖的热阻(junction-to-case thermal resistance, θ_{jc}), 数据中心设施团队则仅聚焦机柜进风温度与环境间的热交换. 二者之间的“冷板-冷却液分配单元-室外冷却塔”链条, 长期为服务器原始设备制造商(original equipment manufacturer, OEM)与基础设施供应商所主导.

然而, 这一链条正被系统性地打通. 英伟达Blackwell GPU的液冷系统, 将PUE从传统风冷模式下的1.5~2.0降至1.15~1.2[14]. 这一变化不仅体现为冷却介质的更替, 更反映在热设计与服务器调度、芯片功率管理的深度联动上. 大型数据中心正通过AI模型对任务负载进行温度预测调度, 在时间与空间维度上均衡分配高热工作负载, 同时结合芯片功耗模型实施动态频率调整, 形成热-功协同调度机制[34]. 这一机制进一步反向驱动芯片架构设计. 传统上, 芯片架构师只需关注电子元器件密度与互连效率; 如今, 数据中心的热管理需求正被纳入前端设计清单[15]: 内存与计算单元采用交错布局以分散热点, 避免热量在局部积聚; 异构核心(高性能大核与能效小核)根据典型负载的产热特性进行冷热分区, 使热流分布更均匀; 片上温度传感器阵列的密度与位置也需针对数据中心的调度策略进行优化配置, 确保温度反馈的实时性与精确性. 这正是热管理作为交叉学科的完整内涵: 其上下游边界已不再清晰, 而是贯穿从EDA工具到冷却塔的整个技术链条[1].

6 新质生产力: 热管理的战略跃迁

2025年党的二十届四中全会后, “新质生产力”成为中国产业升级的核心政策叙事. 热管理正在完成从工程技术支撑到战略基础能力的身份跃迁[1]. 这一跃迁并非单点突破, 而是沿着三重递进逻辑系统性展开: 散热瓶颈从“后端配套”跃升为“前端卡脖子”环节, 迫使热设计从被动响应转向主动牵引芯片架构, 进而驱动热管理从单一散热技术向封装—调度—能源的系统重构演进. 以下分述其内在机理.

6.1 三重跃迁: 从配套到核心

第一重跃迁: 从“辅助环节”到“卡脖子环节”. 2026年, 英伟达下一代Rubin架构GPU的功耗预计将突破2000 W. 芯片性能的提升不再仅依赖于晶体管数量, 而是取决于散热能力的强弱[14]. 散热瓶颈已直接制约计算能力的发展[15]. 这正是新质生产力领域“技术约束倒逼创新”的典型场景.

第二重跃迁: 从“被动应对”到“主动牵引”. 传统产品设计遵循线性逻辑, 即先有芯片, 再匹配散热. 当前, 散热能力已成为决定芯片架构的关键因素[16]. 台积电在晶圆上芯片后接基板(chip-on-wafer-on-substrate, CoWoS)封装环节便与材料供应商联合研发碳化硅中介层; 英伟达下一代GPU平台将液冷接口纳入标准配置并写入参考设计; 数据中心的冷却架构也反向影响服务器主板的布局规范[12]. 热管理不再是后端的补救措施, 而是前端的“定边界”.

第三重跃迁: 从“单点技术”到“系统重构”. 散热效率已不再是唯一的关键绩效指标(key performance indicator, KPI). AI

芯片的热管理需同时考虑封装应力、成本、可量产性及与现有产线的兼容性; 数据中心冷却则需在散热效率、水资源消耗、碳足迹与运营成本之间寻求最优解[12]. 系统层面的最优解优先于局部最优解, 这一要求促使全链条开展协同创新[1].

这三重跃迁并非孤立并列, 而是形成“瓶颈识别—架构牵引—系统重构”的递进闭环, 第三重跃迁的系统级成果(如液冷标准化、PUE体系)又反向缓解第一重跃迁的散热压力.

当散热能力可直接兑换算力成为行业共识, 热管理技术路线的选择将上升为标准之争. 两相浸没与单相冷板之争, 不再仅是工程优化问题, 而将深刻塑造数据中心的物理形态; PUE考核标准的制定权将取决于主流冷却方案的能效基准; 甚至AI芯片的架构设计范式也将被热管理技术重新定义. 掌握了最尖端热管理技术的企业及其所在国家, 将有能力主导下一代算力基础设施的标准制定, 从芯片封装规范到数据中心建筑标准, 从能效评价指标到碳核算方法学. 热管理的终极跃迁, 是从技术跟随者转变为规则制定者.

6.2 人才战: 结构性紧缺的时代

2025年IDTechEx报告指出, 全球具备传热学、流体力学、材料学与电子工程交叉背景的热管理工程师, 其供需缺口正以每年30%的速度扩大[12]. 《中国学科发展战略·电子设备热管理》报告明确指出: “人才短缺已从储备性短缺转变为结构性紧缺. ”[1] 结构性紧缺意味着, 这一问题无法通过单纯增加课程数量或扩大招生规模得以解决. 此类人才需同时掌握声子玻尔兹曼方程与两相流沸腾曲线, 熟悉EDA工具流程与冷却塔选型参数, 并能够在纳米级热斑与兆瓦级热负载之间灵活切换思维尺度, 属于复合型人才群体[1,18].

7 学科的核心挑战

上述讨论表明, 热管理的战略价值已无需赘言, 从AI算力、新能源汽车到商业航天、聚变能源, 热管理无处不在[17]. 然而, 若仅以“有用”定义一门学科, 则失之浅薄. 工程发展史上, 许多曾被视为“有用”的技术方向, 最终在技术迭代中沦为手册里的过时章节[1]. 热管理的独特之处在于: 其底层问题从未被彻底解决, 学科边界仍在持续向外延伸. 这门学科之所以具有持久的生命力, 源于三组核心挑战, 它们共同构成一个闭环的逻辑整体.

第一组挑战: 尺度的巨大跨度. 热管理领域要求研究者在十余个数量级范围内频繁切换认知尺度[18]. 鲜有学科需要研究者如此高频地转换尺度视角, 而这恰是热管理对思维能力的极致训练. 1 nm级: 声子在界面的透射[5]; 1 μm级: 微通道内气泡的成核与脱离[24,25]; 1 mm级: 封装基板的热通孔阵列布局[16]; 1 m级: 浸没冷却槽体内的流动与换热[10]; 1000 km级: 低轨卫星星座的轨道热环境与热控调度.

第二组挑战: 成熟理论与未知机理的并存. 正是这种“尺度的巨大跨度”, 直接导致了宏观理论的成熟与微纳机制的未知并存. 热管理领域的传热学教科书内容(导热、对流、辐射、相变传热)已高度成熟[3]. 任何一本传热学教材都能提供傅里叶定律的数学表达式、努塞尔数的经验关联式以及沸腾曲线的经典形态. 然而, 当研究者面对真实工程问题时, 教科书的内容往往无法给出有效解答[1]. 纳米尺度下, 傅里叶定律失效—教科书没有给出替代方案[5]; 三维堆叠芯片中, 热流路径高度耦合—教科书讲授一维热阻网络, 但考虑微纳尺度效应的三维问题的解析解目前尚不存在[26]; 异质界面处, 理论预测热阻与实测值偏差30%以上—教科书建议“实验测定”[7]; 微重力环境下, 两相浸没冷却的气泡行为完全改变—教科书指引“参考相关文献”[12]. 宏观理论的成熟与微观认知的空白, 正是尺度跨越带来的直接后果.

第三组挑战: 基础与应用的零距离衔接. 面对此类跨越尺度的未知问题, 传统“基础研究→应用转化”的长链条模式已难以满足需求. 在此背景下, 热管理领域形成了基础物理与工程实践的无缝衔接: 研究者可在同一周内, 既推导声子玻尔兹曼方程在超晶格结构中的解析解[20,21], 又将相关模型植入EDA工具的温度感知布局模块[14]; 既能于上午探讨氮化硼/铜复合界面热阻的微观机制[31], 又可在下午前往产线调整热界面材料的涂布工艺参数[9]. 从傅里叶定律到硅片背面蚀刻的微通道, 仅相隔一张实验台的距离[3,8]. 这种“从方程到产线”的短链路, 并非偶然选择, 而是解决跨尺度未知问题的内在要求, 当微观机制尚未被理论完全揭示时, 必须通过基础研究与应用实践的即时反馈填补认知缺口.

这种“基础与应用的零距离”又反过来支撑和扩展着“尺度的巨大跨度”. 正是因为能够在纳米级的声子输运理论与产线的热界面工艺之间无缝切换, 研究者才得以在十余个数量级的尺度间建立有效的知识迁移和模型验证. 基础研究的即时工程化, 使微观新发现能迅速嵌入宏观系统设计中; 而工程实践中的尺度难题, 又持续向基础科学提出新的问题. 三者形成动态闭环: 尺度跨度制造认知鸿沟, 鸿沟催生理论-实践的零距离融合, 零距离融合又助力跨越新的尺度边界.

这三组挑战以闭环逻辑共同定义了热管理的学科特色: 它既扎根于成熟的经典理论, 又不断遭遇前沿的未知问题; 既要求宏观系统思维, 又需要微观物理洞察; 既有明确的工程目标, 又持续向基础科学提出问题. 这正是热管理作为“未完成学科”的根本魅力所在.

8 学科的未来议程

2022年, 中国科学院在《中国学科发展战略·电子设备热管理》发展战略报告中指出: “电子设备热管理学科正处于从‘跟随型’向‘引领型’转变的关键时期. ”[1] 这一判断揭示了热管理学科当前的历史方位. 经过三十年发展, 热管理的学科框架已基本确立, 声子输运理论、界面热阻模型、相变传热机制、液冷技术谱系等核心知识体系已相对成熟. 但这张

学科地图的“海岸线”虽已绘就，“内陆”却仍是亟待探索的空白.

基于上述三组核心挑战，未来研究可收敛为三个需优先突破的关键科学问题:

(1) 跨尺度热输运统一理论: 当器件特征尺寸达到亚10 nm量级时，如何构建突破傅里叶定律局限的热输运理论, 并基于第一性原理精准预测晶体管级别的瞬态温度分布[5,22]? 如何在微通道内实现稳定的流动沸腾传热, 同时避免临界热流密度的触发[24,26]? 虽然玻尔兹曼输运方程可描述非傅里叶效应，但全频率、全波矢求解的计算代价极高，无法直接对接含数十亿晶体管的芯片级EDA仿真. 在亚微米通道内，气泡成核的随机性与表面结构强依赖性，使得临界热流密度的预测模型仍停留在经验关联式层面，无法像经典传热学那样给出普适的解析判据. 这些问题并非短期内可通过工程试错解决，而需要工程热物理、凝聚态物理、计算数学与微电子学的长期协同攻关.

(2) 真实界面热阻的原子级精准预测与调控: 如何在原子尺度层面设计界面，将卡皮查热阻降低至理论极限以下, 并通过声子谱工程实现对界面热导的主动调控[7,9]? 在三维堆叠芯片中，如何规划热流路径，在不挤占互连资源的前提下为每一层芯片开辟专属散热通道[24,26]? 如何将氮化硼纤维、砷化硼单晶等实验室成果转化为可量产、低成本且高可靠的工程材料[31,32]? 虽然石墨烯、碳纳米管等纳米材料具有超高本征热导率，但宏观组装时的接触热阻、与基底界面热阻及长期热循环可靠性仍未突破; 三维堆叠中散热通道与电互连存在版图资源竞争，目前仅能在实验室验证，尚未实现大尺寸量产; 氮化硼、砷化硼等新材料的高质量单晶生长、掺杂工艺及与CMOS产线的兼容性仍是未解难题. 这些瓶颈涉及材料物理、精密制造与封装工艺的深层耦合，非单一技术路线可独立突破, 需作为长期基础研究方向持续投入.

(3) 芯片—数据中心跨尺度热协同优化方法: 如何建立从晶体管到冷却塔的全栈热模型, 实现跨尺度的热-功协同优化[12,14]? 需要重新定义从晶体管到建筑环境的跨尺度热流标准，并建立芯片架构师、热设计工程师与数据中心设施工程师的协同设计范式.

这些问题既见于每本传热学教材的第一章, 也载于DARPA未来十年的项目指南[19], 同时收录于中国科学院的学科发展战略报告[1]，也呈现在台积电、英伟达、微软、谷歌等企业研发白皮书的扉页[14]. 它们是热管理学科留给下一代研究者的“未完成交响曲”.

未来三十年，热管理需完成从“被动响应”到“主动定义”的能力进阶，不仅要攻克现有系统的散热难题，更需通过热设计重新构建芯片架构、数据中心形态乃至能源系统的底层逻辑. 这恰是热管理作为“新质生产力”核心引擎的深层意涵: 它不仅是突破技术瓶颈的破局者，更是重塑产业范式的革新者. 而这场重塑, 方兴未艾.

致谢 本工作得到了国家自然科学基金卓越研究群体项目(T2588301)和重点项目(52036002)与北京市自然科学基金联合重点项目(L257012)的资助. 感谢赵天寿院士等在成稿期间的指导. 本文的插图绘制得到了团队研究生郭欣杰、邹启凡等的帮助, 并使用了AI软件DeepSeek和Gemini.

推荐阅读文献

1 Chinese Academy of Sciences. Disciplinary Development Strategy: Thermal Management of Electronic Equipment (in Chinese). Beijing: Science Press, 2022 [中国科学院. 中国学科发展战略·电子设备热管理. 北京: 科学出版社, 2022]
2 Chen G. Nanoscale Energy Transport and Conversion: A Parallel Treatment of Electrons, Molecules, Phonons, and Photons. Oxford: Oxford University Press, 2005
3 Yang S M, Tao W Q. Heat Transfer (in Chinese). 6th ed. Beijing: Higher Education Press, 2021 [杨世铭, 陶文铨. 传热学. 第6版. 北京: 高等教育出版社, 2021]
4 Guo Z Y. Thermal Fluid Science (in Chinese). Beijing: Tsinghua University Press, 2022 [过增元. 热流体学. 北京: 清华大学出版社, 2022]
5 Cahill D G, Braun P V, Chen G, et al. Nanoscale thermal transport: II. 2003–2012. Appl Phys Rev, 2014, 1: 011305
6 Chen G. Nonclassical heat transfer and recent progress. ASME J Heat Mass Transf, 2024, 147: 032502
7 Swartz E T, Pohl R O. Thermal boundary resistance. Rev Mod Phys, 1989, 61: 605–668
8 Filippou I, Tselepi V, Ellinas K. A review of microfabrication approaches for the development of thin, flattened heat pipes and vapor chambers for passive electronic cooling applications. Micro Nano Eng, 2024, 22: 100235
9 Bashir A, Maqbool M, Khan A Z, et al. Emerging trends and challenges in thermal interface materials: a comprehensive perspective from fundamentals to applications. Mater Sci Eng-R-Rep, 2025, 162: 100789
10 Zou Q F, Liu H, Luo H L, et al. Progress in liquid cooling technology for high-power large-area AI chips (in Chinese). J Refrig, 2026, 47: 20–36 [邹启凡, 刘洪, 罗海亮, 等. 高功率大面积AI芯片液冷技术进展. 制冷学报, 2026, 47: 20–36]

11 Chen Q, Zhou J, Wang J, et al. Review of jet impingement boiling heat transfer: mechanisms, influencing parameters and enhancement strategies. Int J Heat Mass Transfer, 2026, 256: 128133

12 IDTechEx. Thermal Management for Data Centers 2026–2036: Technologies, Markets, and Opportunities. Technical Report. Cambridge: IDTechEx, 2025

13 Chen X G, Hu C D, Liang J, et al. Efficient two-phase liquid cooling and optimization technology on the chip side based on high-power data centers: summary and prospects. Renew Sust Energy Rev, 2025, 223: 115594

14 Siemens EDA. Key thermal advances driving next-gen AI chip design in 2026. (2025-12-15)[2026-04-11]. https://blogs.sw.siemens.com/semiconductor-packaging/2025/12/15/key-thermal-advances-driving-next-gen-ai-chip-design-in-2026/

15 Woon W Y, Kasperovich A, Wen J R, et al. Thermal management materials for 3D-stacked integrated circuits. Nat Rev Electr Eng, 2025, 2: 598–613

16 International Roadmap for Devices and Systems (IRDS). Thermal Management Chapter, 2025 Edition. Technical Report. Piscataway: IEEE, 2025

17 Bai M, Liu N, Liu J, et al. A review of thermal management technologies for electric vehicles. J Energy Storage, 2026, 143: 119658

18 Bar-Cohen A, Holloway C A. Thermal science and engineering: from macro to nano in 200 years. In: Proceedings of the 15th International Heat Transfer Conference (IHTC-15). Kyoto, 2014

19 Joshi Y K, Smith A H, Eden R C, et al. Three decades of thermal management research at DARPA. J Electron Packaging, 2025, 147: 030802

20 Liu T H, Wang T, Zhou J, et al. Thermal-driven multi-carrier transport in electronic and energy materials. Appl Phys Rev, 2025, 12: 031308

21 Gu X, Yang R. Phonon transport and thermal conductivity in two-dimensional materials. Annu Rev Heat Transfer, 2016, 19: 1–65

22 Qian X, Zhou J, Chen G. Phonon-engineered extreme thermal conductivity materials. Nat Mater, 2021, 20: 1188–1202

23 Uchiyama H, Liu T H, Ono T, et al. Quantifying doping-dependent electron-phonon scattering rates in silicon by inelastic x-ray scattering and first-principles lattice dynamics. Phys Rev Mater, 2023, 7: 104601

24 Wen R, Li Q, Wang W, et al. Enhanced bubble nucleation and liquid rewetting for highly efficient boiling heat transfer on two-level hierarchical surfaces with patterned copper nanowire arrays. Nano Energy, 2017, 38: 59–65

25 Liu X, Yang J, Zou Q, et al. Enhancing liquid–vapor phase-change heat transfer with micro/nano-structured surfaces. ACS Nano, 2025, 19: 9513–9589

26 Sharma M K, Ramos-Alvarado B. Thermal management of 3-D heterogeneously integrated microelectronics: challenges and future research directions. Commun Eng, 2026, 5: 28

27 Balandin A A. Thermal properties of graphene and nanostructured carbon materials. Nat Mater, 2011, 10: 569–581

28 Zhang X, Zhou Z, Guo Y, et al. High-coherence parallelization in integrated photonics. Nat Commun, 2024, 15: 7892

29 Huang C, Qian X, Yang R. Thermal conductivity of polymers and polymer nanocomposites. Mater Sci Eng-R-Rep, 2018, 132: 1–22

30 May P W, Zulkharnay R. Diamond thin films: a twenty-first century material. Part 2: a new hope. Phil Trans R Soc A, 2025, 383: 20230382

31 Jiang P, Qian X, Yang R, et al. Anisotropic thermal transport in bulk hexagonal boron nitride. Phys Rev Mater, 2018, 2: 064005

32 Tian F, Song B, Chen X, et al. Unusual high thermal conductivity in boron arsenide bulk crystals. Science, 2018, 361: 582–585

33 Shin J, Gamage G A, Ding Z, et al. High ambipolar mobility in cubic boron arsenide. Science, 2022, 377: 437–440

34 Ranjbar B, Nguyen T D A, Ejlali A, et al. Power-aware runtime scheduler for mixed-criticality systems on multicore platform. IEEE Trans Comput-Aided Des Integr Circuits Syst, 2021, 40: 2009–2023

Summary for “热管理学科前沿: 从微纳传热到AI算力基础设施”

Advances in thermal management: from micro/nanoscale heat transfer to AI computing infrastructure

Rong-Gui Yang[1*] & Xiao-Bo Li[2*]

[1] *Department of Energy and Resources Engineering, College of Engineering, Peking University, Beijing 100871, China*
[2] *School of Energy and Power Engineering, Huazhong University of Science and Technology, Wuhan 430074, China*
* Corresponding authors, E-mail: ronggui@pku.edu.cn; xbli35@hust.edu.cn

Thermal management encompasses the integrated set of technologies and methodologies for actively designing, precisely controlling, and systematically optimizing the generation, transport, distribution, and dissipation of heat within devices and systems, ensuring that operating temperatures remain within safe, efficient, and reliable ranges. From smartphones, robots, and artificial intelligence (AI) chips to data centers, new energy vehicles, and aerospace equipment, thermal management has permeated every arena of modern technology.

However, thermal management is facing unprecedented challenges. The power consumption of a single AI chip has exceeded 1200 W, approaching the power level of a household induction cooker. Power density per rack in data centers has surged from the traditional 10 kW to 250 kW, far exceeding the limits of air cooling. Local heat flux now surpasses 300 W/cm^2, several times that of a nuclear reactor core. When power consumption of a single chip exceeds 1200 W, “cooling capacity is computing power” becomes the new industrial paradigm, and the “Power Wall” has become the primary obstacle hindering the continued progress of Moore’s law.

An even more profound challenge lies in spanning scales: from quantum phonon transport at the 1 nm scale, to bubble nucleation boiling in 1 μm channels, to airflow organization in 1 m cabinets, and thermal control scheduling for satellite constellations spanning 1000 km. Thermal management must establish a unified thermal order across more than ten orders of magnitude in scale. This necessitates deep interdisciplinary collaboration involving condensed matter physics, microelectronics, fluid mechanics, heat transfer, materials science, and computer science.

In 2022, the Chinese Academy of Sciences included “Thermal Management of Electronic Equipment” in its series of disciplinary development strategies, marking the transition of thermal management from an engineering support function to a fundamental field requiring national investment for systematic research. As a key foundation of “new quality productive forces”, thermal management is undergoing an identity transformation from a “supporting role” to a “core constraint”. The discipline has evolved from a passive “back-end engineering” approach to active “front-end thermal design”, where cooling strategies must be integrated from the earliest stages of device architecture and layout planning. Each technological breakthrough in thermal management will open new possibilities for AI computing, clean energy development, and commercial aerospace advancement.

This article presents a comprehensive overview of the discipline, covering its definition and boundaries, historical evolution from macro to nano, fundamental physics including phonon transport and interfacial thermal resistance, key challenges in integrated circuits and AI data centers, and its strategic role as a core enabler of next-generation technologies. It aims to provide a systematic introduction for readers interested in this dynamic and unfinished field.

thermal management, integrated circuits, AI data centers, phonon transport, interfacial thermal resistance, liquid cooling technology

doi: 10.1360/CSB-2026-0245